Elemental estimates, influence, and algorithmic leveraging

Keith Knight

Abstract It is well-known (Subrahmanyam, 1972; Mayo and Gray, 1997) that the ordinary least squares estimate can be expressed as a weighted sum of so-called elemental estimates based on subsets of p observations where p is the dimension of parameter vector. The weights can be viewed as a probability distribution on subsets of size p of the predictors $\{\boldsymbol{x}_i : i = 1, \cdots, n\}$. In this paper, we derive the lower dimensional distributions of this p dimensional distribution and define a measure of potential influence for subsets of observations analogous to the diagonals elements of the "hat" matrix for single observations. This theory is then applied to algorithmic leveraging, which is a method for approximating the ordinary least squares estimates using a particular form of biased subsampling.

1 Introduction

Given observations $\{(\boldsymbol{x}_i, y_i) : i = 1, \cdots, n\}$, we define the ordinary least squares (OLS) estimate $\widehat{\boldsymbol{\beta}}$ as the minimizer of

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2.$$

We are implicitly assuming that $\widehat{\boldsymbol{\beta}}$ estimates a p-dimensional parameter $\boldsymbol{\beta}$ in the model $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i$ $(i = 1, \cdots, n)$ for some errors $\{\varepsilon_i\}$. However, we will not use this assumption in the sequel.

The OLS estimate can be written as a weighted sum of so-called elemental estimates, which are based on subsets of observations of size p. If $s = \{i_1 <$

Keith Knight
University of Toronto, Toronto ON Canada, e-mail: keith@utstat.toronto.edu

$\cdots < i_p\}$ is a subset of $\{1, \cdots, n\}$ then we can define the elemental estimate $\widehat{\boldsymbol{\beta}}_s$ satisfying

$$\boldsymbol{x}_{i_j}^T \widehat{\boldsymbol{\beta}}_s = y_{i_j} \quad \text{for } j = 1, \cdots, p$$

provided that the solution $\widehat{\boldsymbol{\beta}}_s$ exists. Subrahmanyam (1972) showed that the OLS estimate can be written as

$$\widehat{\boldsymbol{\beta}} = \sum_s \frac{|X_s|^2}{\sum\limits_u |X_u|^2} \widehat{\boldsymbol{\beta}}_s$$

where the summation is over all subsets of size p, $|K|$ denotes the determinant of a square matrix K and

$$X_s = \left(\boldsymbol{x}_{i_1} \ \boldsymbol{x}_{i_2} \cdots \boldsymbol{x}_{i_p} \right). \tag{1}$$

Therefore, we can think of the OLS estimate $\widehat{\boldsymbol{\beta}}$ as an expectation of elemental estimates with respect to a particular probability distribution; that is,

$$\widehat{\boldsymbol{\beta}} = E_{\mathcal{P}}(\widehat{\boldsymbol{\beta}}_S)$$

where the random subset S has a probability distribution

$$\mathcal{P}(s) = P(S = s) = \frac{|X_s|^2}{\sum\limits_u |X_u|^2} \tag{2}$$

where X_s is defined in (1). Hoerl and Kennard (1980) note that the OLS estimate can also be expressed as a weighted sum of all OLS estimates based on subsets of $k > p$ observations.

An analogous result holds for weighted least squares (WLS) where we minimize

$$\sum_{i=1}^n w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$$

for some non-negative weights $\{w_i\}$. Again in this case, the WLS estimate $\widetilde{\boldsymbol{\beta}}$ can be written as $\widetilde{\boldsymbol{\beta}} = E(\widehat{\boldsymbol{\beta}}_S)$ where now S has the probability distribution

$$P(S = s) = \frac{|X_s|^2 \prod\limits_{j \in s} w_j}{\sum\limits_u \left\{ |X_u|^2 \prod\limits_{j \in u} w_j \right\}} = \frac{\mathcal{P}(s) \prod\limits_{j \in s} w_j}{\sum\limits_u \left\{ \mathcal{P}(u) \prod\limits_{j \in u} w_j \right\}}.$$

Henceforth, we will focus on the distribution $\mathcal{P}(s)$ defined in (2) for the OLS estimate where the results for the WLS estimate will follow *mutatis mutandis*.

The probability $\mathcal{P}(s)$ defined in (2) describes the weight and therefore the potential influence of a subset s (of size p) on the OLS estimate $\widehat{\boldsymbol{\beta}}$. In particular, greater weight is given to subsets where the vectors $\boldsymbol{x}_{i_1}, \cdots, \boldsymbol{x}_{i_p}$ are more dispersed; for example, if $\boldsymbol{x}_i = (1, x_i)^T$ then $|(\boldsymbol{x}_i\ \boldsymbol{x}_j)|^2 = (x_i - x_j)^2$. We can also use the probability distribution $\mathcal{P}$ to define measures of influence of arbitrary subsets of observations.

In section 2, we will derive the lower dimensional distributions of $\mathcal{P}$ defined in (2) while in section 3, we will discuss the potential influence of a subset of observations.

In situations where n and p are large, the OLS estimate $\widehat{\boldsymbol{\beta}}$ may be difficult to compute in which case one can attempt to approximate $\widehat{\boldsymbol{\beta}}$ by sampling $m \ll n$ observations from $\{(\boldsymbol{x}_i, y_i)\}$ leading to a subsampled estimate $\widehat{\boldsymbol{\beta}}_{ss} = E_{\mathcal{Q}}(\widehat{\boldsymbol{\beta}}_S)$ where S has a distribution $\mathcal{Q}$. The goal here is to find a subsampling scheme so that $\mathcal{Q} \approx \mathcal{P}$ in some sense. This will be explored further in section 4.

2 Lower dimensional distributions of $\mathcal{P}$

The probability distribution $\mathcal{P}$ defined in (2) describes the weight given to each subset of p observations in defining the OLS estimate. It is also of interest to consider the total weight given to subsets of $k < p$ observations. It turns out that these lower dimensional distributions depend on the elements of the so-called "hat" matrix. (The "hat" matrix is the orthogonal projection onto the column space of the matrix X whose rows are $\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_n^T$.)

We start by re-expressing $\mathcal{P}(s)$. Since

$$\sum_u |X_u|^2 = \left| \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \right|$$

(Mayo and Gray, 1997), it follows that

$$\begin{aligned} \mathcal{P}(s) &= \left| X_s^T \left(\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} X_s \right| \\ &= \left| \begin{pmatrix} h_{i_1 i_1} & h_{i_1 i_2} & \cdots & h_{i_1 i_p} \\ \vdots & \vdots & \ddots & \vdots \\ h_{i_p i_1} & h_{i_p i_2} & \cdots & h_{i_p i_p} \end{pmatrix} \right| \end{aligned}$$

where $\{h_{ij} : i, j = 1, \cdots, n\}$ are the elements of the "hat" matrix (Hoaglin and Welsch, 1978):

$$h_{ij} = h_{ji} = \boldsymbol{x}_i^T \left(\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \boldsymbol{x}_j.$$

Henceforth, unless specified otherwise, all probabilities and expected values are based on the probability distribution $\mathcal{P}$.

If S is a random subset of size p drawn from $\{1, \cdots, n\}$ with probability distribution $\mathcal{P}$, it is convenient to describe the distribution of S using the equivalent random vector $\boldsymbol{W} = (W_1, \cdots, W_n)$ where $W_j = I(j \in S)$ and $W_1 + \cdots + W_n = p$. The moment generating function $\varphi(\boldsymbol{t}) = E[\exp(\boldsymbol{t}^T \boldsymbol{W})]$ of $\boldsymbol{W}$ is given by

$$\begin{aligned} \varphi(\boldsymbol{t}) &= \left| \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \right|^{-1} \sum_s |X_s|^2 \left\{ \prod_{i_j \in s} \exp(t_{i_j}) \right\} \\ &= \frac{\left| \sum_{i=1}^n \exp(t_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \right|}{\left| \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^T \right|}. \end{aligned}$$

Thus for $k \leq p$,

$$P(\{i_1, \cdots, i_k\} \subset S) = E\left(\prod_{j=1}^k W_{i_j} \right) = \left. \frac{\partial^k}{\partial t_{i_1} \cdots \partial t_{i_k}} \varphi(\boldsymbol{t}) \right|_{t_1 = t_2 = \cdots = t_n = 0}.$$

The following result gives the lower dimensional distributions of $\mathcal{P}$.

Proposition 1. *Suppose that S has the distribution $\mathcal{P}$ defined in (2). Then for $k \leq p$,*

$$P(\{i_1, \cdots, i_k\} \subset S) = \left| \begin{pmatrix} h_{i_1 i_1} & h_{i_1 i_2} & \cdots & h_{i_1 i_k} \\ \vdots & \vdots & \ddots & \vdots \\ h_{i_k i_1} & h_{i_k i_2} & \cdots & h_{i_k i_k} \end{pmatrix} \right|.$$

Proof. Define the $k \times k$ matrix

$$H_{i_1 \cdots i_k}(\boldsymbol{t}) = \exp(t_{i_1} + \cdots + t_{i_k}) \begin{pmatrix} \boldsymbol{x}_{i_1}^T \\ \vdots \\ \boldsymbol{x}_{i_k}^T \end{pmatrix} \left(\sum_{i=1}^n \exp(t_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} (\boldsymbol{x}_{i_1} \cdots \boldsymbol{x}_{i_k})$$

and define for $1 \leq i, j \leq n$,

$$h_{ij}(\boldsymbol{t}) = \boldsymbol{x}_j^T \left(\sum_{i=1}^n \exp(t_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \boldsymbol{x}_i.$$

It suffices to show that

$$\frac{\partial^k}{\partial t_{i_1} \cdots \partial t_{i_k}} \varphi(\boldsymbol{t}) = \varphi(\boldsymbol{t}) \left| H_{i_1 \cdots i_k}(\boldsymbol{t}) \right|. \tag{3}$$

We will prove (3) by induction using Jacobi's formula (Goldberg, 1972)

$$\frac{d}{dt}|K(t)| = \text{trace}\left(\text{adj}(K(t))\frac{d}{dt}K(t)\right) = |K(t)|\text{trace}\left(K^{-1}(t)\frac{d}{dt}K(t)\right)$$

where $\text{adj}(K(t))$ is adjugate (the transpose of the cofactor matrix) of $K(t)$ as well as the identity

$$\left|\begin{pmatrix} D & \boldsymbol{v} \\ \boldsymbol{v}^T & a \end{pmatrix}\right| = a|D| - \boldsymbol{v}^T \text{adj}(D)\boldsymbol{v} \tag{4}$$

where a is a real number, $\boldsymbol{v}$ a vector of length k, and D a $k \times k$ matrix. For $k = 1$, we have

$$\begin{aligned}
\frac{\partial}{\partial t_{i_1}} \varphi(\boldsymbol{t}) &= \varphi(\boldsymbol{t}) \text{trace}\left\{ \left(\sum_{i=1}^{n} \exp(t_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \exp(t_{i_1}) \boldsymbol{x}_{i_1} \boldsymbol{x}_{i_1}^T \right\} \\
&= \varphi(\boldsymbol{t}) \exp(t_{i_1}) \boldsymbol{x}_{i_1}^T \left(\sum_{i=1}^{n} \exp(t_i) \boldsymbol{x}_i \boldsymbol{x}_i^T \right)^{-1} \boldsymbol{x}_{i_1} \\
&= \varphi(\boldsymbol{t}) \exp(t_{i_1}) h_{i_1 i_1}(\boldsymbol{t}) \\
&= \varphi(\boldsymbol{t}) \left| H_{i_1}(\boldsymbol{t}) \right|.
\end{aligned}$$

Now suppose that (3) holds for some $k < p$ and set $\ell = k + 1$. Then

$$\begin{aligned}
\frac{\partial^\ell}{\partial t_{i_1} \cdots \partial t_{i_\ell}} \varphi(\boldsymbol{t}) &= \frac{\partial}{\partial t_{i_\ell}} \left\{ \varphi(\boldsymbol{t}) \left| H_{i_1 \cdots i_k}(\boldsymbol{t}) \right| \right\} \\
&= \left| H_{i_1 \cdots i_k}(\boldsymbol{t}) \right| \frac{\partial}{\partial t_{i_\ell}} \varphi(\boldsymbol{t}) + \varphi(\boldsymbol{t}) \frac{\partial}{\partial t_{i_\ell}} \left| H_{i_1 \cdots i_k}(\boldsymbol{t}) \right|.
\end{aligned}$$

First,

$$\frac{\partial}{\partial t_{i_\ell}} \varphi(\boldsymbol{t}) = \varphi(\boldsymbol{t}) \left| H_{i_\ell}(\boldsymbol{t}) \right| = \varphi(\boldsymbol{t}) \exp(t_{i_\ell}) h_{i_\ell i_\ell}(\boldsymbol{t}).$$

Second,

$$\varphi(\boldsymbol{t}) \frac{\partial}{\partial t_{i_\ell}} \left| H_{i_1 \cdots i_k}(\boldsymbol{t}) \right| = \varphi(\boldsymbol{t}) \left\{ \text{trace}\left(\text{adj}(H_{i_1 \cdots i_k}(\boldsymbol{t})) \frac{\partial}{\partial t_{i_\ell}} H_{i_1 \cdots i_k}(\boldsymbol{t}) \right) \right\}$$

with

$$\frac{\partial}{\partial t_{i_\ell}} H_{i_1\cdots i_k}(\boldsymbol{t}) = -\exp(t_{i_1}+\cdots+t_{i_k}+t_{i_\ell}) \begin{pmatrix} h_{i_1 i_\ell}(\boldsymbol{t}) \\ \vdots \\ h_{i_k i_\ell}(\boldsymbol{t}) \end{pmatrix} (h_{i_1 i_\ell}(\boldsymbol{t}) \cdots h_{i_k i_\ell}(\boldsymbol{t}))\,.$$

Applying (4) with

$$a = h_{i_\ell i_\ell}(\boldsymbol{t}), \quad D = H_{i_1\cdots i_k}(\boldsymbol{t}), \quad \text{and} \quad \boldsymbol{v} = \begin{pmatrix} h_{i_1 i_\ell}(\boldsymbol{t}) \\ \vdots \\ h_{i_k i_\ell}(\boldsymbol{t}) \end{pmatrix},$$

we get

$$\frac{\partial^\ell}{\partial t_{i_1}\cdots\partial t_{i_\ell}} \varphi(\boldsymbol{t}) = \varphi(\boldsymbol{t})\,|H_{i_1\cdots i_\ell}(\boldsymbol{t})|$$

and the conclusion follows by setting $\boldsymbol{t} = \boldsymbol{0}$.

3 Measuring influence for subsets of observations

The diagonal elements $\{h_{ii}\}$ of the "hat" matrix are commonly used in regression analysis to measure the potential influence of observations (Hoaglin and Welsch, 1978). Similar influence measures for subsets of observations have been proposed; see Chatterjee and Hadi (1986) as well as Nurunnabi *et al.* (2014) for surveys of some of these methods.

From Proposition 1, it follows that $P(W_i = 1) = h_{ii} = E(W_i)$, which suggests that an analogous measure of the influence of a subset of observations whose indices are $i_1, \cdots, i_k$ might be based on the distribution of $W_{i_1}, \cdots, W_{i_k}$.

Suppose that A is a subset of $\{1, \cdots, n\}$ and define

$$N(A) = \sum_{j\in A} W_j. \tag{5}$$

Given that $E(W_i) = h_{ii}$ and and $P(W_i = 1, W_j = 1) = E(W_i W_j) = h_{ii}h_{jj} - h_{ij}^2$ from Proposition 1, it follows that

$$\begin{aligned} E[N(A)] &= \sum_{j\in A} h_{jj} \\ \text{Var}[N(A)] &= \sum_{j\in A} h_{jj} - \sum_{i\in A}\sum_{j\in A} h_{ij}^2. \end{aligned}$$

More generally, the probability distribution of $N(A)$ in (5) can be determined from the probability generating function

$$E\left[t^{N(A)}\right] = \frac{\left|t\sum_{j\in A}\boldsymbol{x}_j\boldsymbol{x}_j^T + \sum_{j\notin A}\boldsymbol{x}_j\boldsymbol{x}_j^T\right|}{\left|\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\right|}.$$

This gives, for example, if $A = \{i_1, \cdots, i_k\}$,

$$\begin{aligned} P\left(N(A)=0\right) &= \frac{\left|\sum_{j\notin A}\boldsymbol{x}_j\boldsymbol{x}_j^T\right|}{\left|\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\right|} \\ &= \left|I - \left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}\sum_{j\in A}\boldsymbol{x}_j\boldsymbol{x}_j^T\right| \\ &= \left|\begin{pmatrix} 1-h_{i_1i_1} & -h_{i_1i_2} & \cdots & -h_{i_1i_k} \\ -h_{i_2i_1} & 1-h_{i_2i_2} & \cdots & -h_{i_2i_k} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{i_ki_1} & -h_{i_ki_2} & \cdots & 1-h_{i_ki_k} \end{pmatrix}\right|. \end{aligned} \tag{6}$$

In the case where $h_{i_1i_1}, \cdots, h_{i_ki_k}$ are uniformly small and $k \ll n$ then

$$P\left(N(A)=0\right) \approx \exp\left(-\sum_{j=1}^{k}h_{i_ji_j} - \frac{1}{2}\sum_{j=1}^{k}\sum_{\ell=1}^{k}h_{i_ji_\ell}^2\right).$$

Also note that (6) can also be computed as

$$\mathcal{P}\left(N(A)=0\right) = \prod_{j=1}^{k}\left\{1 - \boldsymbol{x}_{i_j}^T\left(\sum_{i\in A\setminus\{i_1,\cdots,i_{j-1}\}}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}\boldsymbol{x}_{i_j}\right\}$$

where the quadratic form

$$\boldsymbol{x}_{i_j}^T\left(\sum_{i\in A\setminus\{i_1,\cdots,i_{j-1}\}}\boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}\boldsymbol{x}_{i_j}$$

is a diagonal of the "hat" matrix with observations $i_1, \cdots, i_{j-1}$ deleted.

Suppose that $\widehat{\boldsymbol{\beta}}_A$ is the OLS estimate of $\boldsymbol{\beta}$ based on $\{(\boldsymbol{x}_i, y_i) : i \notin A\}$ and define $\mathcal{P}_A$ to be the probability distribution on subsets S so that $\widehat{\boldsymbol{\beta}}_A = E_{\mathcal{P}_A}(\widehat{\boldsymbol{\beta}}_S)$. If $\mathcal{P}_A$ is close to $\mathcal{P}$ then we would expect $\widehat{\boldsymbol{\beta}}_A$ to be close to $\widehat{\boldsymbol{\beta}}$ —

in other words, the influence of the subset A on estimation of $\boldsymbol{\beta}$ is small. More generally, if we delete the observations in A, we may want to define an estimate based on elemental estimates from subsets s with $s \cap A = \emptyset$ using a different probability distribution $\mathcal{Q}$ (with $\mathcal{Q}(s) = 0$ if $s \cap A \neq \emptyset$) so that

$$\widetilde{\boldsymbol{\beta}}_A = \sum_s \widehat{\boldsymbol{\beta}}_s \mathcal{Q}(s).$$

The following result provides a simple formula the total variation (TV) distance between $\mathcal{P}_A$ and $\mathcal{P}$ as well as giving a condition on $\mathcal{Q}$ that minimizes the TV distance between $\mathcal{Q}$ and $\mathcal{P}$.

Proposition 2. *(a) Define $\mathcal{P}_A(s) = \mathcal{P}(s)/P(N(A) = 0)$ for subsets s with $s \cap A = \emptyset$. Then*

$$d_{tv}\left(\mathcal{P}_A, \mathcal{P}\right) = \sup_B |\mathcal{P}_A(B) - \mathcal{P}(B)| = P\left(N(A) \geq 1\right)$$

where $P\left(N(A) \geq 1\right)$ can be evaluated using (6). (b) Suppose that $\mathcal{Q}$ is a probability distribution on subsets s with $\mathcal{Q}(s) = 0$ if $s \cap A \neq \emptyset$. Then $d_{tv}\left(\mathcal{Q}, \mathcal{P}\right) \geq P\left(N(A) \geq 1\right)$ where the lower bound is attained if $\mathcal{Q}(s) = \lambda(s)\mathcal{P}(s)$ (for $s \cap A = \emptyset$) where $\lambda(s) \geq 1$.

Proof. (a) We can compute the TV distance as

$$d_{tv}\left(\mathcal{P}_A, \mathcal{P}\right) = \frac{1}{2} \sum_s |\mathcal{P}_A(s) - \mathcal{P}(s)|.$$

If $s \cap A = \emptyset$ then

$$\mathcal{P}_A(s) = \frac{\mathcal{P}(s)}{P(N(A) = 0)}$$

with $\mathcal{P}_A(s) = 0$ when $s \cap A \neq \emptyset$. Thus

$$\begin{aligned} d_{tv}\left(\mathcal{P}_A, \mathcal{P}\right) &= \frac{1}{2} \sum_s |\mathcal{P}_A(s) - \mathcal{P}(s)| \\ &= \frac{1}{2} \left\{ \sum_{s \cap A = \emptyset} |\mathcal{P}_A(s) - \mathcal{P}(s)| + \sum_{s \cap A \neq \emptyset} |\mathcal{P}_A(s) - \mathcal{P}(s)| \right\} \\ &= P\left(N(A) \geq 1\right). \end{aligned}$$

(b) For probability distributions $\mathcal{Q}$ concentrated on subsets s satisfying $s \cap A = \emptyset$, we have

$$\sum_{s \cap A \neq \emptyset} |\mathcal{P}_A(s) - \mathcal{P}(s)| = P\left(N(A) \geq 1\right);$$

thus it suffices to minimize

$$\sum_{s \cap A = \emptyset} |\mathcal{Q}(s) - \mathcal{P}(s)|$$

subject to

$$\sum_{s \cap A = \emptyset} \mathcal{Q}(s) = 1.$$

The first order condition implies that the minimizer $\mathcal{Q}^*$ must satisfy $\mathcal{Q}^*(s) \geq \mathcal{P}(s)$ for all s and so $\mathcal{Q}^*(s) = \lambda(s)\mathcal{P}(s)$ where $\lambda(s) \geq 1$ and

$$\sum_{s \cap A = \emptyset} \lambda(s)\mathcal{P}(s) = 1.$$

Now

$$\begin{aligned} d_{tv}\left(\mathcal{Q}^*, \mathcal{P}\right) &= \frac{1}{2}\sum_s |\mathcal{Q}^*(s) - \mathcal{P}(s)| \\ &= \frac{1}{2}\left\{ \sum_{s \cap A = \emptyset} |\mathcal{Q}^*(s) - \mathcal{P}(s)| + \sum_{s \cap A \neq \emptyset} |\mathcal{Q}^*(s) - \mathcal{P}(s)| \right\} \\ &= \frac{1}{2}\left\{ \sum_{s \cap A = \emptyset} (\lambda(s) - 1)\mathcal{P}(s) + \sum_{s \cap A \neq \emptyset} \mathcal{P}(s) \right\} \\ &= P\left(N(A) \geq 1\right), \end{aligned}$$

which completes the proof.

Part (a) of Proposition 2 suggests that $P\left(N(A) \geq 1\right)$ is a natural analogue of the "hat" diagonals for measuring the potential influence of observations with indices in A. More precisely, we can define the leverage $\mathrm{lev}(A)$ of the subset $A = \{i_1, \cdots, i_k\}$ as

$$\mathrm{lev}(A) = P\left(N(A) \geq 1\right) = 1 - \left| \begin{pmatrix} 1 - h_{i_1 i_1} & -h_{i_1 i_2} & \cdots & -h_{i_1 i_k} \\ -h_{i_2 i_1} & 1 - h_{i_2 i_2} & \cdots & -h_{i_2 i_k} \\ \vdots & \vdots & \ddots & \vdots \\ -h_{i_k i_1} & -h_{i_k i_2} & \cdots & 1 - h_{i_k i_k} \end{pmatrix} \right|. \quad (7)$$

As before, if $h_{i_1 i_1}, \cdots, h_{i_k i_k}$ are uniformly small and $k \ll n$ then we can approximate $\mathrm{lev}(A)$ in (7) by

$$\mathrm{lev}(A) \approx 1 - \exp\left(-\sum_{j=1}^{k} h_{i_j i_j} - \frac{1}{2}\sum_{j=1}^{k}\sum_{\ell=1}^{k} h_{i_j i_\ell}^2 \right) \approx \sum_{j=1}^{k} h_{i_j i_j} + \frac{1}{2}\sum_{j=1}^{k}\sum_{\ell=1}^{k} h_{i_j i_\ell}^2.$$

As noted in Draper and John (1981), the matrix in equation (7) as well as its determinant (that is, $1 - \mathrm{lev}(A)$) play a role in a number of diagnostic tests

(for example, those of Andrews and Pregibon (1978) and Cook (1977))for assessing the influence of observations whose indices lie in A; see also Little (1985).

Part (b) of Proposition 2 implies that when $P\left(N(A) \geq 1\right) < 1$, any probability distribution of the form $\mathcal{Q}^*(s) = \lambda(s)\mathcal{P}(s)$ where $\lambda(s) \geq 1$ for $s \cap A = \emptyset$ attains the minimum TV distance to $\mathcal{P}$; this condition is always satsified by $\mathcal{P}_A$. In particular, as $P\left(N(A) \geq 1\right)$ decreases, the family of distributions attaining the minimum TV distance becomes richer. (If we replace the TV distance by the Hellinger distance in part (b) then the minimum is attained uniquely at $\mathcal{P}_A$.)

4 Application: Algorithmic leveraging

In least squares problems where n and p are very large, it is often useful to solve a smaller problem where $m \ll n$ observations are sampled (possibly using some weighting scheme) with $\boldsymbol{\beta}$ estimated using OLS or WLS estimation on the sampled obervations. For example, algorithmic leveraging (Drineas *et al.*, 2011; Ma *et al.*, 2015; Ma and Sun, 2015; Gao, 2016) samples observations using biased sampling where the probability that an observation $(\boldsymbol{x}_i, y_i)$ is sampled is proportional to its leverage h_{ii} or an approximation to h_{ii}; efficient methods for approximating $\{h_{ii}\}$ are discussed in Drineas *et al.* (2012). The sampled observations are then used to estimate $\boldsymbol{\beta}$ using OLS or some form of WLS. In addition, the observations may be also be "pre-conditioned": If $\boldsymbol{y}$ is the vector of responses and X is the $n \times p$ matrix whose i row is $\boldsymbol{x}_i^T$ then we can transform $\boldsymbol{y} \mapsto V\boldsymbol{y}$ and $X \mapsto VX$ for some $n \times n$ matrix; V is chosen so that the "hat" diagonals of VX are less dispersed than those of X.

Suppose that a given subsample does not include observations with indices in A; in the case of leveraging. these observations are more likely have small values of h_{ii} and so $P\left(N(A) \geq 1\right)$ will be smaller than if the observations were sampled using simple random sampling. We now estimate $\boldsymbol{\beta}$ by minimizing

$$\sum_{i \notin A} w_i (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$$

for some weights $\{w_i > 0 : i \notin A\}$. The resulting estimate $\widehat{\boldsymbol{\beta}}_{ss}$ can be written as

$$\widehat{\boldsymbol{\beta}}_{ss} = \sum_{s \cap A = \emptyset} \mathcal{Q}(s) \widehat{\boldsymbol{\beta}}_s$$

where

$$\mathcal{Q}(s) = \frac{\mathcal{P}(s) \prod_{j \in s} w_j}{\sum_{u \cap A = \emptyset} \mathcal{P}(u) \prod_{j \in u} w_j}.$$

From Proposition 2, $\mathcal{Q}$ attains the lower bound on the TV distance to $\mathcal{P}$ if $\mathcal{Q}(s) = \lambda(s)\mathcal{P}(s)$ for some $\lambda(s) \geq 1$ when $s \cap A = \emptyset$; in other words, we require

$$\prod_{j \in s} w_j \geq P\left(N(A) = 0\right) \sum_{u \cap A = \emptyset} \left\{ \frac{\mathcal{P}(u)}{P\left(N(A) = 0\right)} \prod_{j \in u} w_j \right\} \tag{8}$$

for all s with $s \cap A = \emptyset$. The condition (8) is always satisfied if all the weights $\{w_i\}$ are equal, in which case, $\widehat{\boldsymbol{\beta}}_{ss}$ is an OLS estimate. For non-equal weights, the situation becomes more complicated. For example, if $w_i = 1/h_{ii}$ and the variability of $\{h_{ii} : i \notin A\}$ is relatively large then (8) may be violated for some subsets s, particularly when $P\left(N(A) = 0\right)$ is close to 1 (so that the lower bound for the TV distance is close to 0). This observation is consistent with the results in Ma *et al.* (2015) as well as Ma and Sun (2015) where unweighted estimation (setting $w_i = 1$) generally outperforms weighted estimation. Proposition 2 also suggests that it may be worthwhile selecting m observations so as to maximize $P\left(N(A) = 0\right)$ and thereby minimizing the TV distance. This effectively excludes low-leverage observations from the sample, which may not be desirable from a statistical point of view; moreover, determining the exclusion set A will be computationally expensive for large p and n.

To illustrate, we consider a simple linear regression with $\boldsymbol{x}_i^T = (1 \ x_i)$ for $i = 1, \cdots, n = 1000$ where $\{x_i\}$ are drawn from a two-sided Gamma distribution with shape parameter $\alpha = 0.5$; this produces a large number

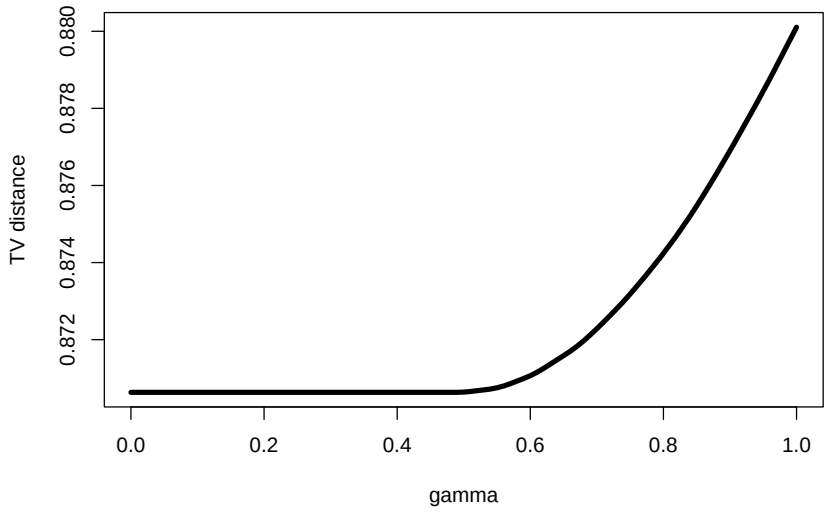

Fig. 1 TV distance as a function of γ for a leverage sample of size $m = 200$.

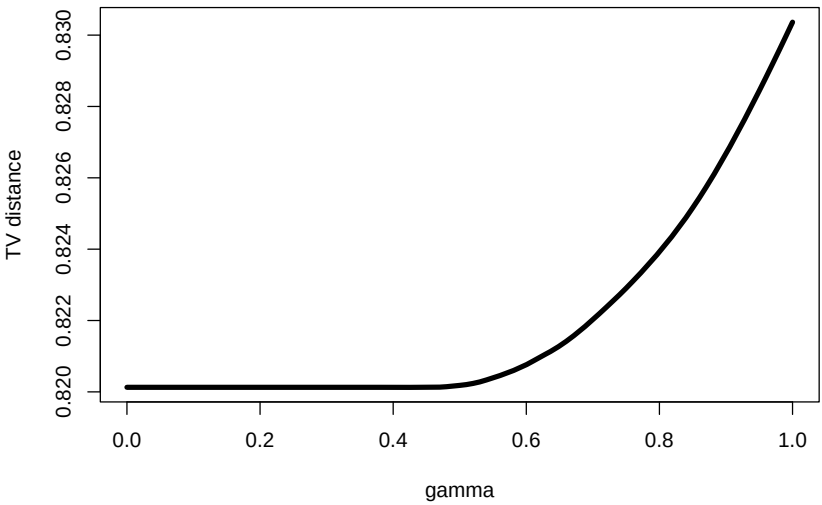

Fig. 2 TV distance as a function of γ for a sample of $m = 200$ where the exclusion set A is chosen to (approximately) maximize $P(N(A) = 0)$.

of both large ($h_{ii} > 4p/n = 0.008$) and small ($h_{ii} \approx 1/n = 0.001$) leverage points. We then draw a sample of 200 (unique) observations using leverage sampling and compute the TV distance for WLS with $w_i = h_{ii}^{-\gamma}$ for $0 \le \gamma \le 1$; a plot of the TV distance as a function of γ is shown in Figure 1. A second sample of 200 (unique) observations is obtaining by excluding a set A of 800 observations to maximize (approximately) $P(N(A) = 0)$; a plot of the TV distance as a function of γ is shown in Figure 2. In both cases, the TV distance is minimized (that is, condition (8) is satisfied) for values of γ between 0 and approximately 0.5 with the minimum TV distance being smaller (0.82 versus 0.87) for the second sample.

References

1. Andrews, D.F., Pregibon, D.:. Finding the outliers that matter. Journal of the Royal Statistical Society. Series B, **40**, 85–93 (1978)
2. Chatterjee, S., Hadi, A.S.: Influential observations, high leverage points, and outliers in linear regression. Statistical Science, **1**, 379–393 (1986)
3. Cook, R.D.: Detection of influential observations in linear regression. Technometrics, **19**, 15–18 (1977)
4. Draper, N.R., John, J.A.: Influential observations and outliers in regression. Technometrics, **23**, 21–26 (1981)
5. Drineas, P., Magdon-Ismail, M., Mahoney, M.W., Woodruff, D.P.: Fast approximation of matrix coherence and statistical leverage. Journal of Machine Learning Research. **1**3, 3475–3506 (2012)

6. Drineas, P., Mahoney, M.W., Muthukrishnan, S., Sarlós, T.: Faster least squares approximation. Numerische Mathematik. **117**, 219–249 (2011)
7. Gao, K.: Statistical inference for algorithmic leveraging. arXiv preprint arXiv:1606.01473 (2016)
8. Golberg, M.A.: The derivative of a determinant. The American Mathematical Monthly. **79**, 1124–1126 (1972)
9. Hoaglin, D.C., Welsch, R.E.: The hat matrix in regression and ANOVA. The American Statistician, **32**, 17–22 (1978)
10. Hoerl, A.E., Kennard, R.W.: M30. A note on least squares estimates. Communications in Statistics–Simulation and Computation, **9**, 315–317 (1980)
11. Little, J.K.: Influence and a quadratic form in the Andrews-Pregibon statistic. Technometrics. **27**, 13–15 (1985)
12. Ma, P., Mahoney, M.W., Yu, B.: A statistical perspective on algorithmic leveraging. Journal of Machine Learning Research. **16**, 861–911 (2015)
13. Ma, P., Sun, X.: Leveraging for big data regression. Wiley Interdisciplinary Reviews: Computational Statistics. **7**, 70–76 (2015)
14. Mayo, M.S., Gray, J.B.: Elemental subsets: the building blocks of regression. The American Statistician. **51**, 122–129 (1997)
15. Nurunnabi, A.A.M., Hadi, A.S.,Imon, A.H.M.R.: Procedures for the identification of multiple influential observations in linear regression. Journal of Applied Statistics. **41**, 1315–1331 (2014)
16. Subrahmanyam, M.: A property of simple least squares estimates. Sankhya, Series B. **34**, 355-356 (1972)